Word Association Thesaurus as a Resource for Extending Semantic Networks

Anna Sinopalnikova[1,2]
[1]*Faculty of Informatics, Masaryk University*
Botanicka 68a
Brno 60200
Czech Republic
[2]*Saint-Petersburg State University*
Universitetskaya 11
Saint-Petersburg 199034
Russia

Pavel Smrz
Faculty of Informatics, Masaryk University;
Botanicka 68a
Brno 60200
Czech Republic

Abstract

The paper reports the on-going research for applying psycholinguistic resources to building and extending semantic networks. We survey different kinds of information that can be extracted from a Word Association Thesaurus (WAT), a resource representing the results of a large-scaled free association test. In addition, we give a comparison of WAT and other language resources (e.g. text corpora, explanatory dictionaries) from the viewpoint of the quality and quantity of semantic information they provide.

1. Introduction

It is generally accepted that we entered the era of semantics (not only linguistic, but semantics in general) and issues of information structuring and retrieval, knowledge representation and understanding are the main directions of nowadays information science.

One of the most popular topics in the areas of modern semantics, information technologies, knowledge representation, natural language processing is the *Semantic Web* (SeW). Usually this term is used to denote the transformation of a present-day World Wide Web into an environment with clear semantics, easily understandable not only by human, but by machines as well. One can consider SeW as being an efficient way of representing data on the WWW, or as of a globally link database.

A special unified format of data presentation and common unified ontologies are recognized as necessary parts of SeW. Although the former already exists in form of RDF and OWL standards, SeW ontological component is still very much in its infancy. There is little consensus about the work on its constructing: its starting point, possible directions and the ways of its accomplishment. Still there is one consideration that is accepted by most people involved: it is unreasonable to build ontologies from scratch. The most likely starting point for SeW building is the efforts to clean-up, refine, standardize and merge the already existing semantic resources: ontologies, lexical databases, semantic networks, etc.

Table 1. Types of existing semantic resources

Corpora	**Dictionaries, thesauri, ontologies, taxonomies**
1. These are primary resources, presenting (more or less) 'raw' data on the language in use. 2. Information is given implicitly. 3. Need special extraction procedures and tools.	1. These are 'derived' resources, presenting explications of some internal knowledge. They are based on primary resources and researcher's intuition. 2. Information is given explicitly.

Roughly speaking, the development of semantic resources follows one of 2 directions: collecting empirical information or creating its logical interpretations (see Table 1).

We will discuss in detail the type of resources that takes intermediate position. It combines the features of primary sources and the structure of derived ones. On the one hand, WAT is close to a corpus because of being a collection of empirical data; on the other hand, it is similar to ontology, because the information is structured in a 'relational' way.

2. Main concepts of psycholinguistics

"We as humans understand the semantics, which means we symbolically represent in some fashion the world, the objects of the world, and the relationships among those objects. We have the semantics of (some part of) the world in our minds; it is very structured and interpreted" [1].

The oldest experimental technique of discovering the way knowledge is structured in the human mind, is the *Word Association (WA) Test*. The first WA test dates back to 1883 [2], slightly modified, it is still in use today. Generally, a list of words *(stimuli)* is given to subjects (either in writing form, or orally), who are asked to respond with the first word that the given word makes them think of *(responses).*

The psycholinguistic term *Association* describes the connection or relation between ideas, concepts, or words, which exists in the human mind and manifests in an above-mentioned way: an appearance of one entity entails the appearance of the other in the mind.

WA tests reveal the respondent's mental model of the world, verbal memories, thought processes, emotional states and personalities. Since 1883 the WA test was applied in various fields of research:

- Reisner [3]: to collect user-oriented retrieval synonyms for IR system
- Rubinoff, Franks and Stone [4]: to provide data about semantic relations between words to be used in building classification schemata
- Palmquist [5]: to expand the search queries
- Pejtersen [6]: to classify paintings
- Ornager [7]: to build image databases etc.

3. Word Association Thesaurus

The results of Word Association Test series carried out with several hundreds stimuli and a few thousand subjects, reported in a form of tables, are known as *Word Association Norms* (WAN). The body of WAN constitutes the list of stimuli, lists of responses and their absolute frequencies for each stimulus word**.** Along with the response distribution, frequency of response is considered to be an essential index, reflecting the strength of semantic relations between words.

Word Association Thesaurus (WAT) is quite similar to WAN, but it excels significantly in size (it includes several thousands of stimuli). Also the procedure of data collection is much more complicated. A small set of stimuli is used as a starting point of the experiment; responses obtained for them are used as stimuli in the next step, the cycle being repeated at least 3 times.

Although WAN are available for hundreds of European, Asian and African languages, WAT were collected only for English and Russian. E.g. Kiss et al [8]: about 54000 words – 1000 subjects; Nelson et al [9]; 75000 responses – 6000 subjects; and Karaulov et al [10]: 23000 words – 1000 subjects.

The advantages of WAT over WAN concern the following points:

- Increasing the number of subjects involved in experiments, we maximize the reliability of the data and the uniformity of responses.
- Increasing the number of words involved in experiments, we approximate the complete presentation of a mental lexicon as a whole.

Therefore, WAT is expected to reflect the basic vocabulary and the basic structure of a particular language (all the relations between words relevant for this particular language system), thus presenting a model of the world of the average native speaker.

4.WAT vs. Corpus

It is unanimously recognized that to build an adequate and reliable semantic network it is not enough to rely upon information produced by 'experts' and stored in traditional resources, whatever advantages for machine usage they offer. One should rather explore the raw data, and extract information from language in its actual (i.e. written and spoken texts), and its potential use (i.e. average speaker's mental lexicon).

Several researchers [11], [12], [13] performed statistical analysis and comparison of such sources of 'raw' data, namely text corpora and word associations, in order to confirm the correlation between frequency of XY co-occurrence in a corpus and the strength of association X-Y in WAT. Those experiments successfully demonstrated that corpora could be used to obtain the same relations between words as WAT. In [14] we made a comparison in the opposite direction, and were to show that a WAT covers more semantic relations than a corpus. For that purpose the Russian WAT [10] and a balanced text corpus of about 16 mln words were used. 6000 'stimulus-response' pairs like *cat – mouse* were extracted from WAT in random order, and then searched in the corpus. The window span was fixed to -10; +10 words.

The most interesting result of our experiment was that about 64% word pairs obtained from subjects do not occur in the corpus (see the first column on Figure1).

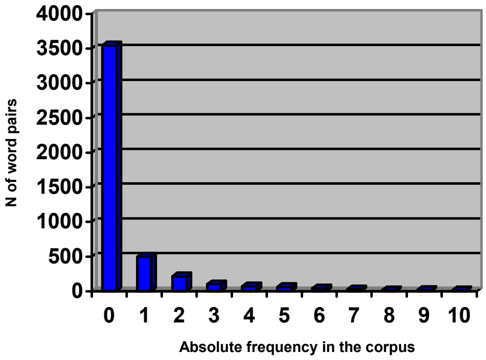

Fig. 1. Overlap between WAT and the corpus.

By excluding all unique associations (that with absolute frequency = 1) from the query list, the proportion of absent pairs may be reduced to 42%, which is still higher than expected. The distribution of the non-unique associations that were not found in the corpus could be seen in Table 2.

Looking for explanation we assumed that paradigmatically related words frequently appear as 'stimulus-respond' and less frequently co-occur in texts. But more detailed observation of the word pairs chosen revealed unexpectedly high ratio of syntagmatic word pairs to be absent. For verbs this number was about 84% of total amount of absent pairs. Whereas paradigmatically related words were regularly presented in the corpus.

.

Table 2. Distribution of word associations that do not occur in the corpus

No of occurrences in the corpus	No of occurrences in WAT	% of all absent assocations
0	2	48
0	3	22
0	4	14
0	5	8
0	6-10	5
0	11-15	<1
0	15-20	<1
0	>20	0

Thus, we conclude that the performed experiment proves the value of WAT as a resource, which could supply the researcher with data otherwise inaccessible.

5. Extracting semantic information from WAT

"Words are really something like condensed actions, situations, and things". [15]

It is obvious that a WAT (or WA test in general) reveals a broad variety of relations stored in the human mind. Some of them are inspired by phonetic or spelling similarity of words, e.g. *know – no, yellow – mellow,* and thus are formal. But the nature of most associations is semantic, 'meaningful' and is caused by relations of objects in reality or respective concepts in the human mind.

In this section we survey different types of semantic data that could be extracted from WAT and applied to semantic network construction:

1) core concepts of the language,
2) semantic primitives,
3) syntagmatic relations between words,
4) paradigmatic relations presented explicitly,
5) domains that are shown,
6) relevance of word senses and relations for native speakers.

5.1 Core concepts

Being a model, WAT reveals the way words are interrelated to each other in the mental lexicon of a native speaker. Experiments [10] proved that each word in our mind is connected to any other word through the path of maximum 6 association links.

However, in WAT there could be observed words that have an above-average number of direct links to other words. I.e. they appear as a response more frequently in the WA tests, e.g Russian *человек, мир*, *дом, любовь, жизнь, есть, думать, жить, идти, большой, хорошо, плохо, нет (не), новый, дерево* etc. *(*295 words with more then 100 relations); English *man, sex, no (not), love, house; work, eat, think, go, live; good, old, small* etc. *(*586 words with more then 100 relations); Czech *člověk, dům, strom; jíst, jít, myslet; moc, starý, velký, bílý, hezký* etc.

The fact is that in every language there is a finite number of such words that appear as responses more frequently then other words. Such sets have several specific features that could be useful:

- they do not change much as the time goes (experiments carried out in 1970-s and 1990-s with Russian subjects gave us the same set of core concepts);
- they do not depend on the starting circumstances, e.g. on words that were chosen as stimulus words;
- children start to learn their native language with these words.

These words determine the fundamental concepts of a particular language system, and thus should be incorporated into ontology as its core components (e.g., SUMO upper concepts [16] or EWN Base Concepts [17]). Representing the most general concepts, these words are associated to most other (more specific) concepts as their superordinate terms (hyponymy relation). We use this information while building wordnets from scratch or controlling already existing ones from the viewpoint of their coverage and consistency.

5.2. Semantic primitives

WAT present information not only about basic concepts for a language as a whole, it also provides a list of basic concepts associated with each separate word. E.g. *Rain – umbrella, drops, wet, storm, thunder, cold, depression…*

We may say that associations reveal semantics of a word (situation) as a list of semantic constituents - separate pieces of information. This data could be very productive while presenting a meaning of a word in terms of minimal semantic constituents. In case of concrete nouns, it is quite easy to present their meaning as finite list of semantic primitives using logic, e.g. *actress – woman + actor.* But semantics of abstract words (verbs, adjectives or nouns denoting complex situation or emotional states) is more difficult to decompose by means of logic and intuition. WAT helps us to solve this problem. E.g. it supplies us with the data that allow to reduce the complex situation of *Depression* to its constituents *sad 7, low 5, black 4, manic 4, sadness 3, bored 3, misery 2, tiredness 2, despair 1, gloom 1, grey 1, hopelessness 1, monotony 1, sick 1, mood 1, nerves 1,* etc., specify its probable causes: *rain 3, guilt 1, pain 1, unemployment 1,* its probable effects: *suicide 1,* its antipodes *elation 3, fun 1, happiness 1* etc., thus placing the concept of depression itself.

5.3. Syntagmatic relations

According to the law of temporal or space contiguity, through life we learn “what goes together” and reproduce it together. Word associations being the “linguistic substitutes for reality” [15] thus reflect the order of events in reality, the way objects are organized in the space, and the way human beings experience them.

Associations by contiguity (that between response and stimulus belonging to different parts of speech) are considered to reflect some syntagmatic relation between respective words. E.g. association *cry – baby* may be treated as a manifestation of relation between verb and its subject, while *take – hand* as a ROLE_INSTRUMENT relation. So, if the stimulus word is a verb, responses are expected to be all its co-occurring words: probable right and left micro-contexts: nouns, adjectives and adverbs that could function in a sentence as its arguments.

In our work we came to the conclusion that a direct incorporation of the word associations into a semantic network could be unreasonable in some cases. As associations reflect instances of relations, there should be a preliminary step of manual analysis and generalization of the data using the hyponymy hierarchy of concepts. E.g. all associations of the same type e.g. *drink – water, beer, milk, ale, Coca-cola, coffee, juice*, etc. found in WAT should be generalized as *drink* ROLE_OBJECT *beverage* relation and in such a form incorporated in the semantic network.

But this does not hold for strongly related words, when we deal with the only possible variant of the relation, e.g. *moo – cow: 70* (ROLE_AGENT), *neigh – horse: 5* (ROLE_AGENT). Such word associations should be incorporated into the semantic networks directly.

5.4. Paradigmatic relations

The law of contiguity could not explain all associations. As experiments by Fillenbaum and Jones [18] shown, it is rare in connected discourse for adjacent words to be from the same part of speech (POS). And the POS category of response is the same as the category of the stimulus word in: 79% associations for nouns, 65% for adjectives, and 43% for verbs. These associations are treated to be caused by the *law of similarity*, and thus pointing to some paradigmatic relation between stimulus and response. E.g. *inanimate – dead: 39* (SYNONYMY), *seek – find: 56* (CAUSE relation), *buy – sell: 56* (CONVERSIVE relation). It is one of the main benefits of WAT that paradigmatic relations are given explicitly as opposed to other sources of empirical data (e.g. text corpora).

WAT turned to be particularly useful for acquiring relations of synonymy and hyponymy. Difference in register, style, or genre prevents co-occurrence of neutral words with 'coloured' ones in text, e.g. *sex – fornicate* (archaic or humorous), *ire* (poetic) *– anger, cowardly – yellow* (slang). Thus, these relations are hardly extractable from texts, but are presented explicitly in WAT. Moreover, this turned to be valid also for some pairs when both synonyms are neutral terms e.g. *astonish – surprise, inanimate – dead, malady – illness* [14].

5.5. Domain information

Apart from the data on conventional set of semantic relations such as synonymy, hyponymy, meronymy etc., WAT provides more subtle information concerning domain structuring of knowledge. E.g., *hospital –> nurse, doctor, pain, ill, injury, load…* This type of data is not so easy to extract from corpora, in explanatory dictionaries it is presented partly (generally covers special terminology only) and mostly based on the lexicographers' intuitions. E.g. *Syringe* – (medicine) *a tube with a nozzle and piston or bulb for sucking in and ejecting liquid in a thin stream* [19]. As opposed to conventional semantic resources, WAT explicitly presents the way common words are grouped together according to the fragments of reality they describe.

Domain relations may be attributed to each concept/word in a semantic network; that give us broader knowledge of the possible contexts for each entry.

These relations are not easily classified, because of the vague distinction of the relations within the situations itself. But according to the frequency we may differentiate the following ones:

- name of domain (situation) – domain member e.g. *hospital – nurse:8, finance – money: 61, football – player:4; marriage – husband 2;*
- participant – participant e.g. *pepper – salt: 58, tamer – lion: 69, needle – thread: 41 mouse – cat: 22;*
- participant – circumstance e.g. *umbrella – rain: 58; actor – stage:23;*
- participant – pointer to its function/role in the situation e.g. *larder – food: 58, envelope – letter: 60, actor – play: 15* etc.

However, it remains arguable whether it is reasonable to differentiate types of domain relations within semantic network, or rather include them as uniform IS_ASSOCIATED_TO relation.

5.6. Applying information from WAT

The above-mentioned methods nave been developed and probed in the process of building specific semantic networks – wordnets, namely RussNet (a wordnet-like database for Russian linking lexical semantics with derivational morphology [20]) and the Czech part of the

BalkaNet project (multilingual wordnet-like network for 5 Balkan languages and Czech [21]).

The experience described in Section 5 was gained in exploring Russian WAT [10]: 8000 stimuli - 23000 words covered – 1000 subjects, and much smaller Czech WAN [22]: 150 stimuli - 4000 words covered – 250 subjects. Also the Edinburgh WAT by Kiss et al [8] has been consulted.

6. Future directions

One of the future directions of our research is the effort to 'mine' common-sense knowledge from WAT. The value and importance of such information in the area of Intelligent Agents have been recognized long time ago [23], but it is still not easily accessible by AI applications. One of the forms of encoding this knowledge is the Minsky's frame [24] or Shank and Abelson's script [25], used to decompose and to represent stereotyped situations or sequences of situations or events. The idea is that things or actions, which are not mentioned explicitly, can be inferred by reference to the script. This enables the agent to "understand" stories and answer questions about them even if the answers are not in the text.

It is one of the interesting features of WAT that it contains the basic information necessary for constructing Minsky's frames. In dealing with this matter we could use the techniques listed in the section 5.2 as a starting point. Certainly, we realize that WAT data could not be automatically converted into the script. However, in combination with other sources of semantic information it could directly form the target descriptions. We plan to test this method within the RussNet and Czech WordNet projects to extend the capacity and applicability of the national wordnets.

7. Conclusions

The advantages of using WAT in constructing semantic networks may be stated as follows:

- *Simplicity* of data acquisition.
- Broad *variety* of semantic information to acquire.

As it was shown, WAT is equal to or excels other sources of semantic information in several respects.

- *Empirical* nature of data extracted (as opposed to theoretical one, cf. conventional ontologies, taxonomies or classification schemes, that supposes the researcher's introspection and intuition to be involved, and hence, leads to over- and under-estimation of the phenomena under consideration).

As it was shown in Section 4, WAT may function as a source of 'raw' data, comparable to a balanced text corpus, and could supply all the necessary empirical information in case of absence of the latter.

- *Probabilistic* nature of data presented (data reflects the relative rather then absolute relevance of semantic relations in each particular case).

References

[1] Daconta, M. C., Obrst, L. J., Smith, K.T. The Semantic Web. Wiley Publishing, Indiana, 2003.

[2] Galton, F. Psychometric Experiments. In: Brain, 2, 1880. pp. 149-162.

[3] Reisner, P. Evaluation of a "growing" thesaurus. Yorktown Heights, IBM Watson Research Center, 1966.

[4] Rubinoff, M. A rapid procedure for launching a microthesaurus. In IEEE, 9 (1), 1966. pp. 8-14.

[5] Palmquist, R.A., Balakrishnan, B. Using a continuous word association test to enhance a user's description of an information need. A quasi-experimental study. In: Proceedings of the 51st ASIS meeting. Atlanta, Georgia, October 23-27, 1988. Medford, NJ, 1988. pp. 160-163.

[6] Pejtersen, A.M. Interfaces based on associative semantics for browsing in information retrieval. Roskilde, Riso Laboratory, 1991.

[7] Ornager, S. Image retrieval: theoretical analysis and empirical user studies on accessing information in images. In: In: Proceedings of the 60th ASIS annual meeting. Washington DC, November 1-6, 1997. Medford, NJ, 1997. pp. 202-2014.

[8] Kiss, G.R., Armstrong, G., Milroy, R. The Associative Thesaurus of English. Edinburg, 1972.

[9] Nelson, D.L., McEvoy, C.L., Schreiber, T.A. The University of South Florida Word Association, Rhyme, and Word fragment norms. 1998. http://www.usf.edu/FreeAssociation/.

[10] Karaulov, Ju.N., Cherkasova, G. A., Ufimtseva, N.V., Sorokin, Ju. A., Tarasov, E.F. Russian Associative Thesaurus. Moscow, 1994-1998.

[11] Church, K. W., Hanks, P. Word Association Norms, Mutual Information, and Lexicography. Computational Linguistics 16 (1). MIT Press, 1990. pp.22-29.

[12] Wettler, M., Rapp R. Computation of Word Associations Based on the Co-Occurrences of Words in Large Corpora. In Proceedings of the 1st Workshop on Very Large Corpora: Academic and Industrial Perspectives. Columbus, Ohio, 1993. pp. 84-93.

[13] Willners, C. Antonyms in Context: A Corpus-based Semantic Analysis of Swedish Descriptive Adjectives. PhD thesis: Lund University Press, 2001.

[14] Sinopalnikova A., Smrz P. Word Association Norms as a Unique Supplement of Traditional Language Resources. In: Proceeding of LREC 2004. Lisboa, 2004 (to be published)

[15] Jung, C. G. The Association Method. In: American Journal of Psychology, 31, 1910. pp. 219-269.

[16] Niles, I., and Pease, A. Origins of the Standard Upper Merged Ontology: A Proposal for the IEEE Standard Upper Ontology. In: Working Notes of the IJCAI-2001 Workshop on the IEEE Standard Upper Ontology, Seattle, Washington, August 6, 2001.

[17] Vossen, P., ed. EuroWordNet: A Multilingual Database with Lexical Semantic Network. Dodrecht, Kluwer , 1998.

[18] Fillenbaum, S., and Jones, L. V. Grammatical Contingencies in Word Association. In: Journal of Verbal Learning and Verbal Behavior, 4, 1965. pp. 248-255.

[19] New Oxford Dictionary of English. Oxford University Press, 1998.

[20] Azarova, I., Mitrofanova, O., Sinopalnikova, A., Yavorskaya, M., Oparin, I. RussNet: Building a Lexical Database for the Russian Language. In: Proceedings: Workshop on Wordnet Structures and Standardisation and How this affect Wordnet Applications and Evaluation. Las Palmas, 2002. pp. 60-64 http://www.phil.pu.ru/depts/12/RN

[21] Stamou, S. et al. BalkaNet: A Multilingual Semantic Network for the Balkan Languages. In: Proceedings of the 1st International Global WordNet Conference. January 21-25, 2002. Mysore. Mysore, India, 2002. pp. 12-14. http://www.ceid.upatras.gr/Balkanet/

[22] Novák, Z. Volné slovní párové asociace v češtině. Praha, 1988.

[23] Lenat, D. B. and Guha, R. V. Building Large Knowledge Based Systems. Reading, Massachusetts: Addison Wesley, 1990. http://www.cyc.com/

[24] Minsky, M. A Framework for Representing Knowledge. In: The Psychology of Computer, 1975.

[25] Shank, R., and Abelson, R. Scripts, Plans, Goals and Understanding. L. Erlbaum Associates, 1977.